Гимназија „Јован Јовановић Змај"
Нови Сад

Матурски рад из вероватноће
Пробабилистичка класификација

Професор ментор:
др Данијела Рајтер-Ћирић

Ученик:
Павле Тепавчевић, IV-12

Нови Сад, мај 2020.

Садржај

1 Увод

Наука и индустрија константно постављају проблеме које статистика треба да реши. У почетку су ово били пољопривредни и индустријски експерименти који су били релативно малог обима. Доласком рачунара и информационог доба комплексност и величина ових проблема су се знатно повећали. Изазови из подручја складиштења, организације и претраге података довели су до настанка нових области науке. Огромне количине података су генерисане, а статистичарев посао је да из њих извуче шаблоне и склоности, и да разуме шта нам подаци говоре. Ово називамо *учење из података*.

Изазови учења из података су довели до наглог развоја статистичких наука. С обзиром на то да рачунање игра веома битну улогу, не треба да нас чуди чињеница да су развоју много допринели истраживачи из других области попут информатике и инжињерства.

Проблеми учења које ми разматрамо се могу поделити у две категорије: надгледане и ненадгледане. У надгледаном учењу циљ је предвидети исход на основу многобројних параметара уноса; у ненадгледаном учењу, не мери се исход, већ нам је циљ да опишемо везе између низа уноса.

Статистичко учење игра важну улогу у многим областима науке, финансија и индустрије. Ево неких примера оваквих проблема:

- Предвидети да ли ће пацијент, који је хоспитализован због инфаркта, поново имати срчани удар. Предвиђање се прави на основу исхране и клиничких мерења тог пацијента.

- Предвидети цену акција за 6 месеци на основу учинка компаније и економских података.

- Читање руком написаних бројева са дигиталне слике.

- Проценити количину глукозе у крви дијабетичара на основу спектра инфрацрвене апсорпције крви те особе.

- Препознавање фактора ризика за добијање карцинома простате на основу клиничких и демографских фактора.

Наука о учењу игра кључну улогу у областима статистике, истраживању података и вештачке интелигенције.

У општем случају имамо исход, углавном квантитативан (нпр. цена неке акције) или категоричан (нпр. има инфаркт/нема инфаркт), који желимо да предвидимо на основу скупа карактеристика (нпр. исхрана и клиничка мерења). Имамо скуп података за обуку у коме проучавамо исход у зависности од карактеристике скупа објеката. Користећи ове податке правимо модел за предвиђање, тј. *ученика*, који нам омогућава да предвидимо исходе нових објеката које до сада нисмо видели. Добар ученик је онај ученик који прецизно предвиђа такве исходе.

2 Надгледано учење

Променљиве који се мере се називају *уноси*. Оне имају утицај на један или више *излаза*. Циљ је искористити уносе да би се предвидели излази. Ово се назива *надгледано учење*.

2.1 Врсте променљивих и нотација

Излази се разликују. У примеру предвиђања количне глукозе у крви излаз је *квантитативан*, дакле неки излази су већи а неки мањи, а они који су сличне вредности су слични и у својој природи. У примеру где се чита руком написан број са дигиталне слике излаз је један од 10 различитих класа цифара: $\mathcal{G} = \{0, 1, ..., 9\}$. Такав излаз се назива *квалитативан* излаз.

За обе врсте излаза има смисла користити уносе да бисмо их предвидели. На пример, користећи нека специфична јучерашња и данашња атмосферска мерења желимо да предвидимо сутрашњи ниво озона. Ако су нам дате вредности количине светлости за сваки пиксел дигитлане слике ручно написаног броја, можемо да предвидимо којој класи ће та слика припадати.

С обзиром на то да имамо два типа излаза постоје и два типа процеса за предвиђање тих излаза: *регресија* је процес предвиђања квантитативног излаза, а *класификација* квалитативног излаза. Видећемо да ова два задатка имају много тога сличног, а оба се могу посматрати као задатак апроксимације функције, тако да су у том погледу еквивалентни.

Квалитативне вредности су углавном представљене нумерички помоћу шифара. Најједноставнији пример је када постоје само две класе односно категорије, као на пример успех и неуспех. Они су обично представљени као бинарна цифра 0 или 1, или као бројеви -1 и 1. Када постоје више од две категорије, постоји неколико алтернатива. Најчешће коришћено шифровање се врши помоћу измишљене променљиве. Овом методом се квалитативна променљива K-тог реда представља помоћу вектора од K јединица и нула, где је у сваком тренутку тачно једна једнака јединици.

Углавном унос означавамо симболом X. Ако је X вектор његове компоненте се означавају са X_j. Квантитативни излази се означавају са Y, а квалитативни са G. Користимо велика слова попут X, Y и G када говоримо о општим аспектима променљивих. Посматране вредности се бележе у променљивама написаним малим словима, дакле i-та посматрана вредност X-а се пише x_i (где је x_i скалар или вектор). Матрице су представљене болдованим великим словима; нпр. скуп од N уноса вектора x_i дужине p бисмо представили $N \times p$ матрицом $\mathbf{X}$. Вектори неће бити представљени болдованим словима, осим када имају N компонената; ово је договор да бисмо разликовали векторе уноса (i-тог посматрања) x_i који су дужине p од вектора x_j дужине N који се састоји од свих посматрања на променљивој X_j.

У овом тренутку можемо широко формулисати задатак учења на следећи начин: за дату вредност вектора уноса X, направити добро предвиђање исхода Y, које означавамо са $\hat{Y}$.

За двокласне излазе G, један начин за њихово решавање јесте да их бинарно шифрујемо као Y и да их посматрамо као квантитативан излаз. Предвиђања за Y ће се углавном налазити на интервалу $[0, 1]$, а ми ћемо доделити вредности класи $\hat{G}$ у зависности од тога да ли важи $\hat{y} > 0.5$ или не. Овај приступ се генерализује за квалитативне излазе K-тог нивоа.

Да бисмо поставили правила предвиђања потребни су нам подаци, неретко много њих. Сходно томе претпоставимо да имамо доступан скуп мерења (x_i, y_i) или $(x_i, g_i), i = 1, ..., N$, познатији као *подаци за обуку*, помоћу којих правимо правила предвиђања.

2.2 Два једноставна приступа предвиђању: метод најмањих квадрата и метод најближих суседа

У овом поглављу представљамо две једноставне али моћне методе предвиђања: линеарни модел који одговара методи најмањих квадрата и методи k-најближих-суседа. Линеарни

модел прави велике претпоставке о структури и доводи до стабилних али понекад непрецизних предвиђања. Метод k-најближих-суседа прави веома благе структуралне претпоставке: његова предвиђања су углавном прецизна али могу бити нестабилна.

2.2.1 Линеарни модели и метод најмањих квадрата

Линеарни модел је био ослонац статистике претходних 40 година и и даље је један од најбитнијих алата. За дат вектор уноса $X^T = (X_1, X_2, ..., X_p)$, ми предвиђамо излаз Y помоћу модела

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

Термин $\hat{\beta}_0$ је грешка у претпоставци у алгоритму за учење. Углавном је згодно укључити константу 1 у X, укључити $\hat{\beta}_0$ у вектор коефицијената $\hat{\beta}$ и онда написати линеарни модел у векторском облику:

$$\hat{Y} = X^T \hat{\beta}$$

где X^T означава транспоновану матрицу или вектор.

Како правимо линеарни модел на основу скупа података за обуку? Постоје многи начини да се то уради, а најпопуларнији јесте *метод најмањих квадрата*. У овом приступу, коефицијенте β бирамо тако да бисмо минимизовали суму квадрата

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2.$$

$RSS(\beta)$ је квадратна функција параметара што значи да њен минимум увек постоји. Решење се најлакше уочава када се ова једначина напише помоћу матрица. Дакле важи,

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta),$$

где је $\mathbf{X}$ матрица $N \times p$ где је свака врста заправо улазни вектор, а $\mathbf{y}$ је вектор излаза скупа за обуку дужине N. Ако узмемо извод по β ове једначине добијамо:

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0.$$

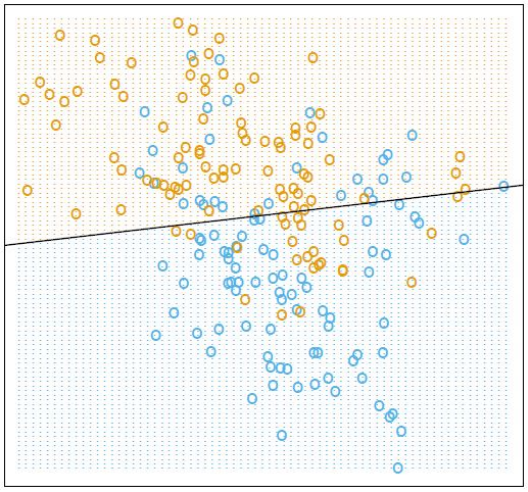
Ако је $\mathbf{X}^T\mathbf{X}$ инвертибилна матрица онда постоји јединствено решење:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Посматрајмо пример линеарног модела у контексту класификације. Слика 1 нам показује дијаграм расејања података за обуку на пару уноса X_1 и X_2. Класа излаза G има две вредности: плаво и наранџасто. Обе класе имају по 100 тачака. Модел линеарне регресије је направљен на основу ових података, са одговором Y шифрованим тако да 0 представља плаву а 1 наранџасту боју. Направљене вредности $\hat{Y}$ су претворене у вредности класе $\hat{G}$ на основу правила:

$$\hat{G} = \begin{cases} \text{наранџаста}, & \text{ако } \hat{Y} > 0.5, \\ \text{плава}, & \text{ако } \hat{Y} \leq 0.5. \end{cases}$$

Можемо приметити неколико погрешних класификација на обе стране границе. Да ли је наш линеарни модел превише строг или су такве грешке неизбежне? Касније ћемо видети



Слика 1: *Пример класификације у две димензије. Класе су шифроване као бинарне променљиве (плаво=0, наранџасто = 1), а онда одређене линеарном регресијом. Линија представља границу дефинисану са* $x^T\hat{\beta} = 0.5$. *Нарандџасто шрафирано подручје представља део улазног простора уноса класификованих као наранџасто, док је плави део класификован као плаво.*

да је линеарна граница најбоља коју можемо да констуишемо, и да је наша процена скоро оптимална.

Посматрајмо сада другу врсту класификације и регресије.

2.2.2 Методи најближих суседа

Методи најближих суседа посматрају једну или више најближих тачака тачки x у пољу уноса и на тај начин формирају $\hat{Y}$. Конкретно, k-најближих суседа одређује $\hat{Y}$ на следећи начин:

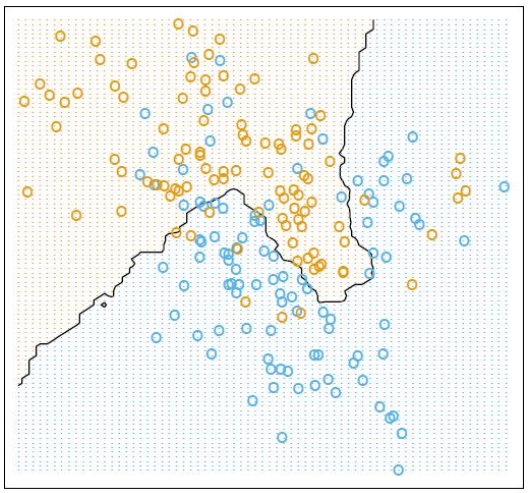
$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

где је $N_k(x)$ област уноса x дефинисан са k најближих тачака x_i у скупу за обуку. Близина подразумева метрику, за коју ми сматрамо да је Еуклидова. Дакле, другим речима, нађемо k најближих тачака x-а (у скупу за обуку) и узмемо просек њихових исхода.

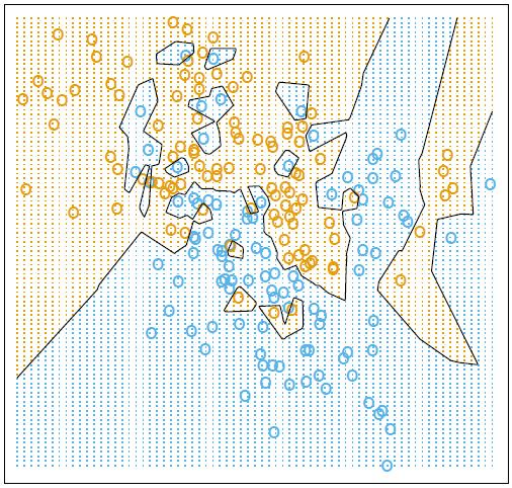
На Слици 2 користимо исте податке за обуку као и на Слици 1, али сада користимо метод 15 најближих суседа. Према томе $\hat{Y}$ представља пропорцију наранџастих тачака у области одређеној са 15 најближих суседа и због тога додељујемо класу наранџасто у $\hat{G}$ ако је $\hat{Y} > 0.5$. Обојени делови простора означавају све тачке у простору уноса које се класификују као плаве односно наранџасте користећи горе описану методу. Видимо да је граница која раздваја плаве од наранџастих регија много неправилнија, и одговара локалним групама где једна класа доминира.

На Слици 3 можемо да видимо како изгледа граница која је направљена помоћу методе 1 најближег суседа. Приметимо да је граница још неправилнија него пре.

На Слици 2 можемо да видимо да је много мање тачака смештено у погрешну класу. Ово не мора да значи да је овај метод бољи. На Слици 3 све тачке су смештене у праве



Слика 2: *Исти пример класификације у две димензије као на Слици 1. Класе су кодиране бинарним променљивама и онда одређене методом 15 најближих суседа.*



Слика 3: *Исти пример класификације у две димензије као на Слици 1. Класе су шифроване као бинарне променљиве, а онде одређене методом 1 најближег суседа.*

класе, али је очигледно да овај начин прављења границе није оптималан.

2.2.3 Од методе најмањих квадрата до методе најближих суседа

Линеарна граница одлуке направљена помоћу методе најмањих квадрата је веома глатка. Она много зависи од претпоставке да је линеарна граница прикладна граница одлуке за посматрани проблем. Другим речима, има ниску *варијансу*, а потенцијално високу *пристрасност*.

У другу руку, метода k најближих суседа се не ослања ни на какве претпоставке о подацима за обуку, односно може да се прилагоди свакој ситуацији. Међутим, сваки део простора који је ограничен границом одлуке зависи од неколико тачака и њихових позиција,

и према томе је нестабилна. Другим речима, има високу *варијансу* и ниску *пристрасност*.

За обе методе постоје ситуације у којима оне раде најбоље. Велики део најпопуларнијих техника које се данас користе су варијације ове две методе. Заправо, метода 1 најближег суседа, најједноставнија од свих, обухвата велики проценат тржишта за мали број димензија. Следећа листа представља неке начине којима се могу оптимизовати ове једноставне методе:

- Кернел методе користе тежине које се глатко смањују до нуле сразмерно дистанци од посматране тачке, уместо тежина 0/1 које су коришћене у методи k најближих суседа.
- У високо-димензионим просторима тежине су модификоване тако да би нагласиле неке променљиве више од других.
- Модели неуронских мрежа се састоје из много трансформисаних линеарних модела.

2.3 Локалне методе у високим димензијама

Чини се да бисмо са довољно великим скупом за обуку могли увек да апроксимирамо границу одлуке методом k најближих суседа. Овај метод не функционише у високим димензијама због феномена познатог под именом *клетва димензионалности* [2]. Постоје многа испољавања овог проблема, а један од најједноставнијих је следећи: Да бисмо у 10 димензионом простору обухватили свега 1% података потребно је да обухватимо више од 63% домена сваког уноса.

2.4 Кернел методе и локална регресија

Може се сматрати да ови методи експлицитно дају процене регресионе функције односно очекивања тако што дефинишу које тачке су суседи једне другима. Он је дефинисан *кернел функцијом* $K_\lambda(x_0, x)$ која додељује вредности тежина тачкама x у околини око x_0. На пример у Гаусовој кернел функцији тежине су одређене на основу функције нормалне расподеле:

$$K_\lambda(x_0, x) = \frac{1}{\lambda} \exp\left[-\frac{\|x - x_0\|^2}{2\lambda}\right]$$

и додељују тежине тачкама тако да оне експоненцијално опадају у односу на квадрат њихове Еуклидске раздаљине од x_0. Параметар λ одговара варијанси нормалне расподеле. Најједноставнији облик кернелове процене јесте Надараја-Ватсон тежински просек:

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}.$$

Метод најближих суседа се може сматрати као специфичан пример кернел метода. Заиста

$$K_k(x, x_0) = I(\|x - x_0\| \leq \|x_{(k)} - x_0\|),$$

где $x_{(k)}$ представља k-ту најкраћу раздаљину од x_0, а $I(S)$ индикатор скупа S.

Ове методе наравно морају бити модификоване у високим димензијама да би се избегла клетва димензионалности.

2.5 Детаљније о надгледаном учењу

Претпоставимо, једноставности ради, да су грешке адитивне и да је модел $Y = f(X) + \epsilon$ разумна претпоставка. Надгледано учење покушава да открије f. Посматрајмо уносе и излазе који су нам дати на почетку и направимо скуп за обуку $\mathcal{T} = (x_i, y_i), i = 1, ..., N$. Вредности уноса x_i су такође унети у систем, познатији као алгоритам за учење (који је углавном компјутерски програм), који, у односу на уносе, прави излазе $\hat{f}(x_i)$. Алгоритам учења има својство да може да мења функцију $\hat{f}$ у зависности од вредности разлике $y_i - \hat{f}(x_i)$. Овај процес се назива *учење примером*. Циљ нам је да вредности вештачких ($\hat{f}(x_i)$) и правих (y_i) излаза након овог процеса буду довољно близу тако да можемо користити ову функцију за предвиђање излаза за будуће уносе чији прави излаз не знамо.

3 Линеарне методе регресије

3.1 Увод

Линеарни модел регресије претпоставља да је функција регресије линеарна у уносима $X_1, ..., X_p$. Линеарни модели су развијени пре компјутерске ере, али су и дан данас веома корисни за учење и чак коришћење. Једноставни су и неретко обезбеђују адекватан опис како уноси утичу на излаз. Што се предвиђања тиче они понекад могу надмашити модерније нелинеарне моделе, поготово у ситуацијама са малим бројем података за обуку. Такође, линеарне методе могу бити примењене на трансформисане уносе што додатно проширује њихову област употребе.

У овом поглављу описујемо линеарне методе регресије, док у следећем дискутујемо о линеарним методима класификације. На неким темама ћемо се задржати дуже јер сматрамо да је темељно разумевање линеарних метода неопходно за разумевање оних нелинеарних. Штавише, многе нелинеарне технике су директна генерализација линеарних метода које ћемо разматрати овде.

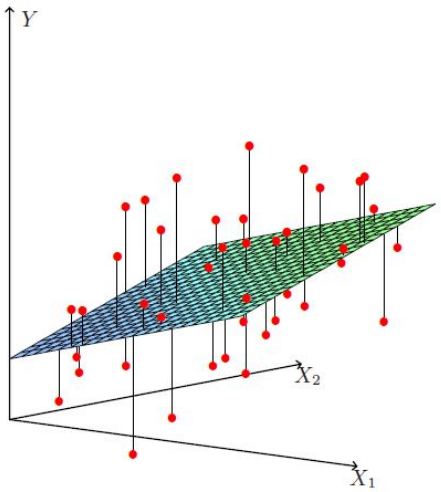
3.2 Модели линеарне регресије и метод најмањих квадрата

Као што смо навели у другом поглављу, имамо вектор уноса $X^T = (X_1, X_2, ..., X_p)$, а желимо да предвидимо квантитативан излаз Y. Модел линеарне регресије има следећи облик:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j.$$

Линеарни модел или претпоставља да је функција регресије $E(Y \mid X)$ линеарна, или да је линеарни модел разумна апроксимација. Вредности β_j су непознати параметри или коефицијенти, а променљиве X_j могу бити:

- квантитативне
- облици квантитативних уноса попут логаритама, корена или квадрата
- полиноми
- нумерички шифровани нивои квалитативних вредности. На пример, ако је G унос са пет нивоа, можемо направити $X_j, j = 1, ..., 5$ таквих да $X_j = I(G = j)$. Заједно ова група X_j-ова представља G, пошто је у суми $\sum_{j=1}^{5} X_j \beta_j$, један од X_j-ова једнак јединици, док су остали једнаки нули.



Слика 4: *Линеарни метод најмањих квадрата фитован за* $X \in \mathbb{R}^2$. *Тражимо линеарну фукцију* X*-а која минимизира суму квадрата остатка* Y*-а.*

- Узајамно дејство између променљивих, на пример $X_3 = X_1 \cdot X_2$.

Независно од облика X_j, модел је линеаран.

Углавном имамо скуп за обуку $(x_1, y_1), ..., (x_N, y_N)$ уз помоћ кога процењујемо параметре β. Сваки $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ је вектор мерења у i-том мерењу. Најпопуларнији метод процене јесте метод најмањих квадрата, у којем бирамо коефицијенте $\beta^T = (\beta_0, \beta_1, ..., \beta_p)^T$ тако да минимизујемо следећу суму:

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \textstyle\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \end{aligned}$$

Са статистичке тачке гледишта, овај критеријум има смисла у случају да су подаци за обуку (x_i, y_i) случајно изабрани из њихове популације. Слика 4 геометријски илуструје методу најмањих квадрата у $\mathbb{R}^{p+1}$ простору.

3.3 Гаус-Марковљева теорема

Један од најпознатијих резултата у статистици тврди да процена параметара β помоћу методе најмањих квадрата има најмању варијансу од свих непристрасних процена.

4 Линеарне методе класификације

4.1 Увод

У овом поглављу посвећујемо пажњу линеарним методама класификације. С обзиром на то да предиктор $G(x)$ узима вредности из дискретног скупа $\mathcal{G}$, ми увек можемо поделити простор улаза у колекцију области означених вредностима које може имати предиктор. Границе ових области могу бити грубе или глатке у зависности од предикционе функције.

Процедуре код којих су границе одлуке линеарне називамо линеарним методама класификације.

Постоји више различитих начина на које можемо да нађемо линеарне границе одлуке. Претпоставимо да постоји K класа, погодности ради означених са $1, 2, ..., K$, и да су њихови одговарајући линеарни модели $\hat{f}_k(x) = \beta_{k0} + \beta_k^T x$. Граница одлуке између класа k и l је скуп тачака за које важи $\hat{f}_k(x) = \hat{f}_l(x)$, тј. $\{x : \hat{\beta}_{k0} - \hat{\beta}_{l0} + (\hat{\beta}_k - \hat{\beta}_l)^T x = 0\}$, што представља праву, раван или хиперраван у зависности од броја димензија у којој правимо класификацију. С обзиром на то да је ово тачно за било који пар класа то значи да је простор подељен у области помоћу хиперпланарних граница. Овај начин регресије припада класи метода које користе функције $\delta_k(x)$ за сваку класу и онда x убаце у класу са највећом вредношћу те функције. Методе које рачунају вероватноће $\Pr(G = k|X = x)$ такође припадају овој класи метода. Очигледно, ако су $\delta_k(x)$ или $\Pr(G = k|X = x)$ линеарне по x онда је и граница одлуке такође линеарна.

Заправо, довољно је да је нека монотона трансформација $\delta_k(x)$ или $\Pr(G = k|X = x)$ линеарна да би и граница одлуке била линеарна. На пример, ако имамо две класе, популаран начин за израчунавање вероватноћа јесте:

$$\Pr(G = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)},$$

$$\Pr(G = 2|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}.$$

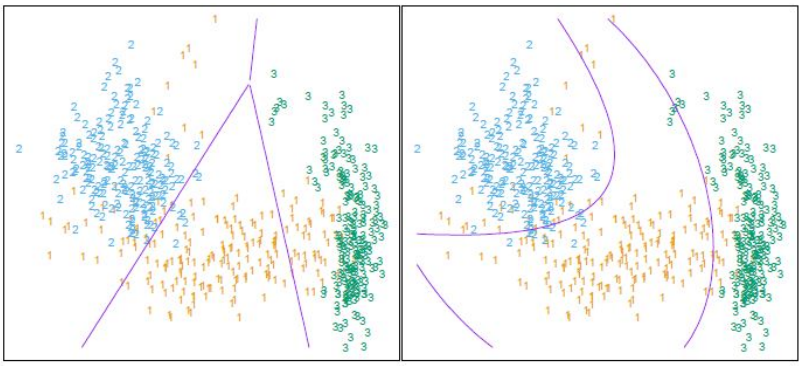
У овом случају је та монотона трансформација заправо логаритамска трансформација $\log[p/(1-p)]$. Заиста функција:

$$\log \frac{\Pr(G = 1|X = x)}{\Pr(G = 2|X = x)} = \beta_0 + \beta^T x.$$

јесте линеарна. За тачке које се налазе на граници одлуке важи да је вредност логаритамске трансформације једнака нули, а та хиперраван, која заправо представља ту границу, је дефинисана са $\{x|\beta_0 + \beta^T x = 0\}$. Постоје две популарне али и различите методе које користе логаритамске трансформације: линеарна дискриминантна анализа (ЛДА) и линеарна логистичка регресија. Њихова суштинска разлика јесте начин на који се прави линеарна функција која одговара скупу података за обуку.

Директнији приступ би био да експлицитно моделујемо границе између класа тако да буду линеарне. Ако посматрамо проблем са две класе у p-димензионом простору, то значи да је наш посао да моделујемо хиперраван. Такође постоје и две методе које експлицитно конструишу те хиперравни. Први алгоритам проналази граничне хиперравни ако оне постоје. Други алгоритам проналази оптималну граничну хиперраван ако она постоји, а ако не постоји онда проналази хиперраван која минимизира неку меру преклапања скупа за обуку.

Иако ћемо већину времена посветити линеарним границама одлуке, постоји знатан потенцијал за оптимизацију. На пример, можемо проширити унешене податке $X_1, ..., X_p$ тако што бисмо додали њихове квадрате и векторске производе $X_1^2, X_2^2, ..., X_1X_2, ...$ Линеарне функције у увећаном простору би се свеле на квадратне функције у првобитном простору. Дакле на овај начин бисмо уопштили линеарне границе одлуке у квадратне границе одлуке. Слика 5 приказује ову идеју. Подаци су исти, на левом графику је приказана линеарна граница одлуке у дводимензионом простору, док се на десном графику прво конструише линеарна граница одлуке у проширеном, петодимензионом, простору који се касније враћа у дводимензиони простор и тако добијамо квадратну границу одлуке.



Слика 5: *Леви график приказује неке податке из три класе раздељене линеарном границом одлуке нађеном линеарном дискриминантном анализом. Десни график приказује квадратне границе одлуке. Оне су добијене првобитним наласком линеарних граница у петодимензионом простору* $X_1, X_2, X_1X_2, X_1^2, X_2^2$. *Линеарне неједнакости у овом простору су квадратног облика у првобитном.*

4.2 Линеарна дискриминантна анализа

Теорија одлуке тврди да су нам потребне вредности $\Pr(G|X)$ да бисмо направили оптималну класификацију. Претпоставимо да је $f_k(x)$ густина X-а у класи $G = k$, и нека је π_k априорна вероватноћа класе k, с тим што важи $\sum_{k=1}^{K} \pi_k = 1$. Применом Бајесове теореме добијамо:

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Претпоставимо да моделујемо сваку густину класе као мултиваријантну нормалну расподелу:

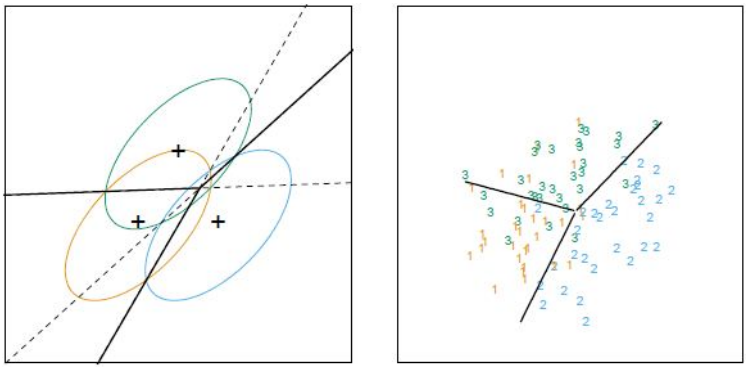
$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}.$$

Линеарна дискриминантна анализа (ЛДА) настаје у специјалном случају када претпоставимо да класе имају исту коваријансну матрицу, тј. меру јачине везе између промене две променљиве $\forall k \Sigma_k = \Sigma$. Када поредимо две класе k и l довољно је гледати логаритам њиховог односа. Можемо видети да је функција

$$\log \frac{\Pr(G = k|X = x)}{\Pr(G = l|X = x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T\Sigma^{-1}(\mu_k - \mu_l) + x^T\Sigma^{-1}(\mu_k - \mu_l),$$

линеарна по x. Ова линеарна логаритамска функција нам говори да је граница одлуке између класа k и l - скуп тачака за које важи $\Pr(G = k|X = x) = \Pr(G = l|X = x)$ - линеарна по x у p димензија. С обзиром на то да је ово тачно за сваке две класе то значи да су све границе линеарне. Ако поделимо простор $\mathbb{R}^p$ у области које класификујемо као класа 1, класа 2,... онда ће ове области бити подељене хиперравнима. Слика 6 нам показује идеализован пример са 3 класе и $p = 2$.

Овде су подаци настали од три нормалне расподеле са истом коваријансном матрицом. Додали смо на слику контуре које одговарају простору где се очекује да се налази 95%



Слика 6: *Лева слика показује три нормалне расподеле са једнаком коваријансном матрицом а различитим очекивањем. Додате су контуре које одговарају простору где се очекује да се налази 95% података. Границе одлуке су приказане између сваке две класе. На десној слици је приказан узорак од 30 тачака које су добијене нормалном расподелом, као и границе одлуке добијене линеарном дискриминантном анализом.*

података. Такође смо додали центроиде класа. Приметимо да границе одлуке нису вертикалне бисектрисе дужи којима су спојени центроиди. Ово би био случај ако би коваријансне матрице Σ биле сферичне јер би у том случају контура која одговара простору где је очекивано да се налази 95% података била у облику круга, а граница одлуке између два круга је симетрала дужи која спаја њихове центре.

У пракси ми не знамо параметре нормалне расподеле што значи да ћемо морати да их процењујемо на основу скупа података за обуку.

На Слици 6 (десна слика) се види граница одлуке направљена на основу узорка од по 30 података за сваку од 3 нормалне расподеле. Слика 5 је такође пример линеарне дискриминантне анализе, али овде класе немају нормалну расподелу.

Иако у случају када имамо више од две класе ЛДА није исто што и линеарна регресија ипак веза између њих може бити успостављена.

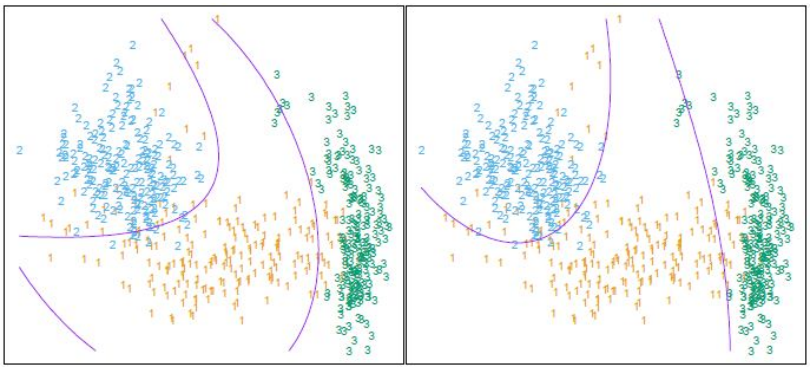
У генералном дискриминантном проблему, ако Σ_k нису међусобно једнаке, онда немамо погодна потирања приликом израчунавања $f_k(x)$, односно квадратни делови остају. Онда добијамо *квадратне дискриминантне фунцкије*,

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log(\pi_k).$$

Граница одлуке између свака два пара класа k и l су описана квадратном једначином $\{x : \delta_k(x) = \delta_l(x)\}$.

Слика 7 приказује пример са Слике 5 где су границе одлуке апроксимиране квадратним једначинама. Овде илуструјемо два начина на који се оне могу поставити тако да одговарају подацима. Десна граница користи КДА, као што смо описали овде, док лева користи ЛДА у повећаном петодимензионом простору. Разлике су мале али се углавном користи КДА.

И ЛДА и КДА функционишу веома добро на великом и разноликом скупу проблема класификације. На пример, у STATLOG пројекту [5] ЛДА је била међу три најбоља класификатора за 7 од 22 скупа података, док је КДА била међу најбоља три за 4 скупа података. Обе методе су широко коришћене и читаве књиге су посвећене њима. Чини се да без обзира на то колико егзотичне методе постојале, да је увек добро имати на располагању ове две алатке.



Слика 7: *Две методе постављања квадратних граница. Леви график показује квадратну границу одлуке за податке са Слике 5 (добијену користећи ЛДА у петодимензионом простору $X_1, X_2, X_1X_2, X_1^2, X_2^2$). Десни график показује квадратну границу нађену помоћу КДА. Разлика је мала што углавном и јесте случај.*

4.3 Логистичка регресија

Логистички регресиони модел је настао из потребе за моделом који израчунава вредности вероватноћа свих K класа користећи функције које су линеарне по x, а да притом припадају интервалу $[0, 1]$ и да је њихов збир једнак јединици. Модел је направљен на следећи начин:

$$\log \frac{\Pr(G = 1|X = x)}{\Pr(G = K|X = x)} = \beta_{10} + \beta_1^T x$$

$$\log \frac{\Pr(G = 2|X = x)}{\Pr(G = K|X = x)} = \beta_{20} + \beta_2^T x$$

$$\vdots$$

$$\log \frac{\Pr(G = K - 1|X = x)}{\Pr(G = K|X = x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$$

Модел је одређен помоћу $K - 1$ услова, односно логаритамских трансформација. Иако користи последњу класу као имениоца, процене су еквивалентне за било који други избор. Тривијално је доказати да је сума

$$\Pr(G = k|X = x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}, k = 1, ..., K - 1,$$

$$\Pr(G = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)},$$

свих вероватноћа једнака један. Да бисмо нагласили зависност модела од скупа параметара $\theta = \{\beta_{10}, \beta_1^T, ..., \beta_{(K-1)0}, \beta_{K-1}^T\}$, означићемо вероватноће са: $\Pr(G = k|X = x) = p_k(x; \theta)$.

За $K = 2$, модел је једноставан јер имамо свега једну линеарну функцију. Он има широку примену у биомедицини где су резултати бинарни (тј. имамо две класе). На пример, пацијент може да преживи или да умре, да има болест срца или не...

4.3.1 Пример: Болести срца у Јужноафричкој Републици

Овде представљамо анализу бинарних података да бисмо илустровали традиционалну примену логистичког регресионог модела. Подаци на Слици 8 су подскуп основног истраживања о коронарном фактору ризика, које се спровело у 3 рурална дела Западног Кејпа [3]. Циљ истраживања је био да се установи фактор ризика болести срца у региону са великом учесталошћу истих. Подаци представљају белце старости између 15 и 64 година, док су одговори присуство или одсуство срчаног удара за време истраживања. Имамо 160 случајева и 302 контроле у скупу података за обуку.

Правимо модел логистичке регресије. Резиме, који се налази на Слици 8, укључује резултат Z за сваки од коефицијената у моделу. Незнатан Z резултат значи да се коефицијент може избацити из модела. Негативан Z резултат значи да би било пожељно да се коефицијент избаци.

	Coefficient	Std. Error	Z Score
(Intercept)	−4.130	0.964	−4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	−1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

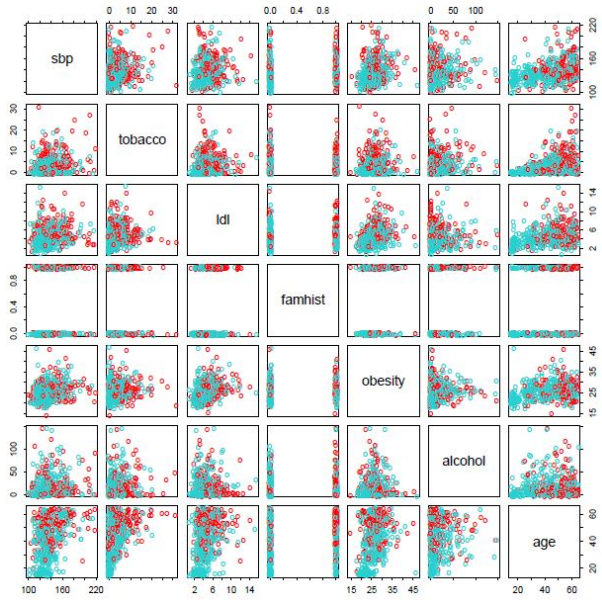
Слика 8: *Резултати добијени логистичким регресионим моделом који одговара подацима срчаних обољења у Јужноафричкој Републици.*

Ако не посматрамо ову табелу пажљиво може се десити да направимо превиде. На пример, ови подаци нам говоре да су горњи крвни притисак (sbp) и гојазност (obesity) небитни. Ова збрка се десила зато што постоји повезаност између скупа предиктора. Они су сами по себи битни, и вредност резултата Z им јесте позитивна. Међутим, у присуству многих других података који су у узајамној вези са њима они више нису потребни, штавише могу имати негативну вредност.

Сада аналитичар може извршити селекцију променљивих, односно избацити непотребне. Један начин да се ово уради јесте да се избаци онај коефицијент који је најмање битан и онда понови поступак, дакле поново направи модел и израчунају вредности које се налазе на Слици 8. Ово се понавља све док се више ни један коефицијент не може избацити из модела. Овим поступком смо добили модел који се налази на Слици 10.

Постоји бољи али и временски захтевнији приступ а то је да направимо модел за сваки од случајева када избацимо једну променљиву и онда за сваки од њих анализирамо девијацију и тако одлучимо коју променљиву да одстранимо.

Поставља се логично питање. Шта нам уопште представљају ови бројеви? На пример, шта нам представља вредност 0.081 на слици 10? "Tobacco", односно дуван, означава колико је особа потрошила дувана за цео свој живот, са просеком од 1.0кг за особе које су се контролисале и 4.1кг за случајеве код којих су потврђена болест срца. Ово нам говори да ако потрошимо 1кг дувана више за време нашег живота, то значи да су нам шансе да имамо срчани удар веће за $\exp(0.081) = 1.084$ односно 8.4%.



Слика 9: *Дијаграм расејања матрице података о болестима срца у Јужноафричкој Републици. Сваки график приказује два фактора ризика. Случајеви и контроле су означене различитим бојама (црвено су случајеви). Променљива (famhist) нам говори да ли се болест јављаја у фамилији и предсављена је бинарно (да или не).*

	Coefficient	Std. Error	Z score
(Intercept)	−4.204	0.498	−8.45
tobacco	0.081	0.026	3.16
ldl	0.168	0.054	3.09
famhist	0.924	0.223	4.14
age	0.044	0.010	4.52

Слика 10: *Резултати добијени логистичким регресионим моделом који одговара подацима срчаних обољења у Јужноафричкој Републици.*

4.4 Логистичка регресија или ЛДА?

У поглављу 4.2 дошли смо до закључка да је логаритам односа вероватноћа за класу k и класу K заправо линеарна функција по x:

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=l|X=x)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k+\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l) + x^T\Sigma^{-1}(\mu_k-\mu_l) = \alpha_{k0} + \alpha_k^T x.$$

Ова линеарност је последица претпоставке да се густина података понаша као нормална расподела и да је коваријансна матрица иста за све класе. У поглављу 4.3 видимо да линеарни логистички метод има линеарне логаритме (ово знамо јер смо га ми тако конструисали):

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} = \beta_{k0} + \beta_k^T x.$$

Чини се да су модели исти. Истина је да имају идентичну форму, али се разлика крије у

томе на који начин су линеарни коефицијенти процењени. Модел направљен логистичком регресијом је општији, тј. прави мање претпоставки.

Шта добијамо додатним ограничавањем модела? Ако се ослонимо на додатне претпоставке онда самим тим имамо више информација о самим параметрима и због тога их можемо боље проценити (нижа варијанса). Ако је $f_k(x)$ баш нормална расподела онда бисмо игноришући ту чињеницу изгубили око 30% ефикасности [4]. Отприлике говорећи, са 30% више података бисмо на овај начин процене параметара добили једнако добру процену као да смо користили чињеницу да је расподела нормална на првобитном скупу података за обуку.

Очигледно је да и подаци који нису сврстани ни у једну класу нам такође доносе неке информације попут коваријансне матрице расподеле или позиције центроида. Често је скупо генерисати класе, али подаци који нису класификовани су прилично јефтини. Ослањајући се на претпоставке о моделу, као што смо то овде радили, можемо да користимо обе врсте информација.

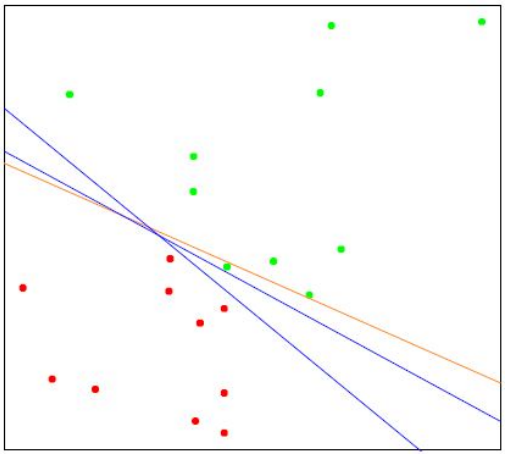
У пракси ове претпоставке никада нису у потпуности тачне, стога људи углавном сматрају да је логистичка регресија сигурнија од ЛДА јер се ослања на мање претпоставки о моделу.

4.5 Хиперравни раздвајања

У овом поглављу наводимо класификаторе који користе хиперравни раздвајања. Они конструишу линеарне границе одлуке које деле податке у класе на најбољи могући начин.

Слика 11 показује 20 тачака у две класе у простору $\mathbb{R}^2$. Ови подаци могу бити подељени линеарном границом. На слици се налазе две плаве линије које представљају једне од бесконачно много хиперравни раздвајања. Наранџаста линија је добијена методом најмањих квадрата. Она је дефинисана са:

$$\{x : \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 0\}.$$



Слика 11: *Пример две класе подељене једном хиперравни. Наранџаста линија је добијена методом најмањих квадрата, која у овом примеру лоше класификује једну тачку из података за обуку. Такође су приказане две плаве раздвајајуће хиперравни.*

У овом случају метод најмањих квадрата прави једну грешку односно не раздваја класе на савршен начин. Иста граница би била добијена линеарном дискриминантном анализом.

4.5.1 Розенблатов алгоритам учења - перцептрон

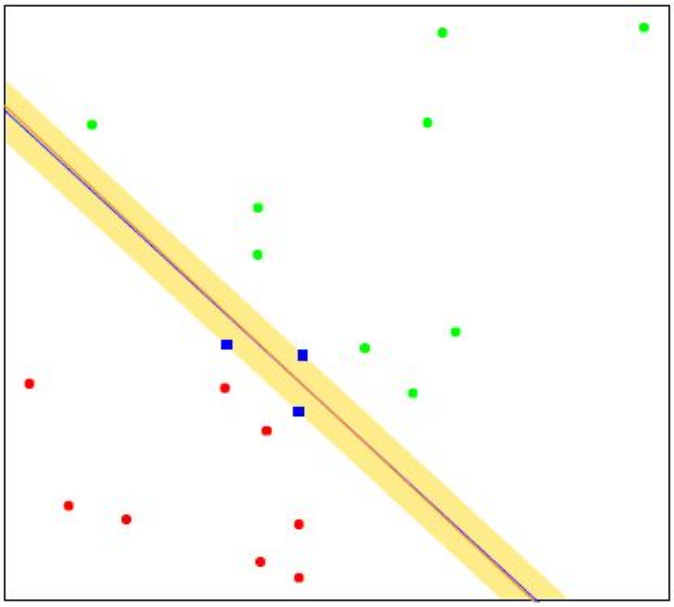
Овај алгоритам учења проналази раздвајајуће хиперравни тако што минимизује раздаљину лоше класификованих тачака од границе одлуке. У случају када имамо две класе ако је $y_i = 1$ лоше класификовано, онда важи $x_i^T \beta + \beta_0 < 0$, а ако је $y_i = -1$ лоше класификовано онда важи $x_i^T \beta + \beta_0 > 0$. Циљ је минимизовати

$$D(\beta, \beta_0) = -\sum_{i \in \mathcal{M}} y_i (x_i^T \beta + \beta_0),$$

где $\mathcal{M}$ представља скуп индекса података који су лоше класификовани. Ова вредност је ненегативна и пропорционална укупној раздаљни лоше класификованих тачака од границе одлуке која је дефинисана са $\beta^T x + \beta_0 = 0$.

4.5.2 Оптималне хиперравни раздвајања

Оптимална хипер раван раздвајања дели две класе и максимизује раздаљину до најближе тачке, независно којој класи она припада. Не само да ово пружа јединствено решење за раздвајајућу хипер раван већ и максимизује разлику између две класе што доводи до бољих класификација података.



Слика 12: *Овде су приказани исти подаци као и на Слици 11. Шрафирани део приказује највећу раздаљину која се може направити између две класе. Овај (шрафирани) простор је одређен трима тачкама. Плава линија, која представља оптималну хипер раван раздвајања, га дели на два једнака дела. На слици се такође налази и граница одлуке добијена логистичком регресијом (црвена линија), која је веома близу оптималне равни раздвајања.*

5 Непознати појмови

Априорна вероватноћа - вероватноћа догађаја пре узимања у обзир релевантна мерења. Ова вероватноћа може бити одређена субјективним осећајем или претпоставком да су сви догађаји једнако вероватни

Варијанса - математичко очекивање одступања случајне променљиве од њене средње вредности

Девијација - одступање од неке вредности

Монотона трансформација - начин пресликавања једног скупа бројева у други тако да поредак вредности остане непромењен

Мултиваријантна нормална расподела - мултиваријантно уопштавање једнодимензионе нормалне расподеле. Описује заједничку расподелу случајног вектора чији су уноси међусобно независне универијантне нормалне случајне променљиве

Стабилност предвиђања - количина промене предвиђања кад се подаци за обуку мало измене

Предиктор - функција која користи скуп коефицијената (независне променљиве) да предвиди исход зависне променљиве

Прецизност предвиђања - проценат предвиђања који је модел добро предвидео

Пристрасност - систематско (уграђено) одступање које чини да све вредности буду погрешне за одређени износ

Теорија одлуке - математичко проучавање стратегија за оптимално одлучивање међу опцијама које укључују различите ризике или очекивања добитка или губитка, зависно од исхода

Центроид - средњи аритметички положај тачака

Литература

[1] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). *The Elements of Statistical Learning*

[2] Bellman, R. 1961. Curse of dimensionality. *Adaptive control processes: a guided tour. Princeton, NJ.*

[3] Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities, *South African Medical Journal* **64**: 430-436.

[4] Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis, *Journal of the American Statistical Association* **70**: 892-898.

[5] Michie, D., Spiegelhalter, D. and Taylor, C. (eds) (1994). *Machine Learning, Neural and Statistical Classification*, Ellis Horwood Series in Artificial Intelligence, Ellis Horwood.

6 Биографија матуранта

Моје име је Павле Тепавчевић. Рођен сам 17. 6. 2001. у Новом Саду, од мајке Милијане и оца Мирослава.

Као веома млад сам се бавио спортом и представљавши Основну школу „Прва војвођанска бригада" сам освојио сребрну медаљу на Олимпијским спортским играма ученика Републике Србије. Након тога сам се спортом наставио бавити рекреативно а своју пажњу сам усмерио према образовању. Након завршених шест разреда основне школе прешао сам у Гимназију „Јован Јовановић Змај" где сам и завршио основно образовање носећи диплому „Вук Караџић". Своје школовање сам наставио у Гимназији „Јован Јовановић Змај" уписавши смер за обдарене ученике у математичкој гимназији који ове године и завршавам. Планирам да упишем Природно-математички факултет у Новом Саду на смеру математика.

Посебно сам поносан на своју такмичарску каријеру. Освојио сам бар 9 медаља, 123 дипломе, од којих 34 на државном и једну на међународном нивоу, а као највећи успех бих издвојио 7 узастопних учешћа на Државним такмичењима из математике, диплому победника међународног математичког турнира градова и III награду на Државном такмичењу из физике за ученике првог разреда која ми је обезбедила право на стипендију за изузетно надарене ученике и студенте коју и дан данас примам.

Датум предаје матурског рада:_____

Комисија:

Председник:_____

Испитивач:_____

Члан:_____

Коментар:

Датум одбране:_____ Оцена_____( )